Model-based co-clustering of multivariate time-dependent data

14th International Conference of the ERCIM WG - CMStatistics 2021

Alessandro Casa

Joint work with: C. Bouveyron, E. Erosheva & G. Menardi

School of Mathematics and Statistics

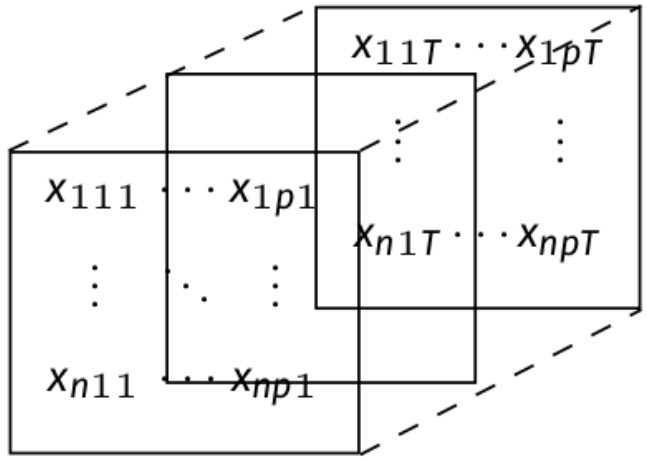
University College Dublin

alessandro.casa@ucd.ie

18th December 2021

Framework

- Increasingly common multivariate time-dependent data can be arranged according to a three-way structure



- Three modes of the data introduce three different challenges
  - rows $\longleftrightarrow$ heterogeneous units
  - columns $\longleftrightarrow$ dependent variables
  - layers $\longleftrightarrow$ correlated occasions

Aim

Extract information and unveil
parsimonious patterns from such data

Co-clustering in a nutshell

Idea

Co-clustering tools may be helpful by summarizing the data into homogeneous blocks

- Given $\mathbf{X} \in \mathbb{R}^{n \times p}$, the Latent Block Model (LBM) is written as

$$p(\mathbf{X}; \Theta) = \sum_{z \in Z} \sum_{w \in W} \prod_{ik} \pi_k^{z_{ik}} \prod_{jl} \rho_l^{w_{jl}} \prod_{ijkl} p(x_{ij}; \theta_{kl})^{z_{ik} w_{jl}}$$

  - $\Theta = (\pi_k, \rho_l, \theta_{kl})_{1 \leq k \leq K, 1 \leq l \leq L}$, with K and L the number of row and column clusters
  - $\mathbf{z} = (z_{ik})_{1 \leq i \leq n, 1 \leq k \leq K}$ and $\mathbf{w} = (w_{jl})_{1 \leq j \leq p, 1 \leq l \leq L}$ denote the subject and variable cluster memberships
  - $(x_{ij} | z_{ik} = 1, w_{jl} = 1) \sim p(\cdot; \theta_{kl})$

Time-dependent Latent Block Model

- In this framework $\mathbf{X} = \{x_{ij}(\mathbf{t}_i)\}_{1 \leq i \leq n, 1 \leq j \leq p}$, $\mathbf{t}_i = (t_{i,1}, \ldots, t_{i,m_i})$
  $\longrightarrow$ each single cell is a curve
- We resort to the Shape Invariant Model (SIM) defined as

$$(x_{ij}(t)|z_{ik} = 1, w_{jl} = 1) = \alpha_{ij,1}^{kl} + e^{\alpha_{ij,2}^{kl}} m(t - \alpha_{ij,3}^{kl}; \beta_{kl}) + \varepsilon_{ij}(t)$$

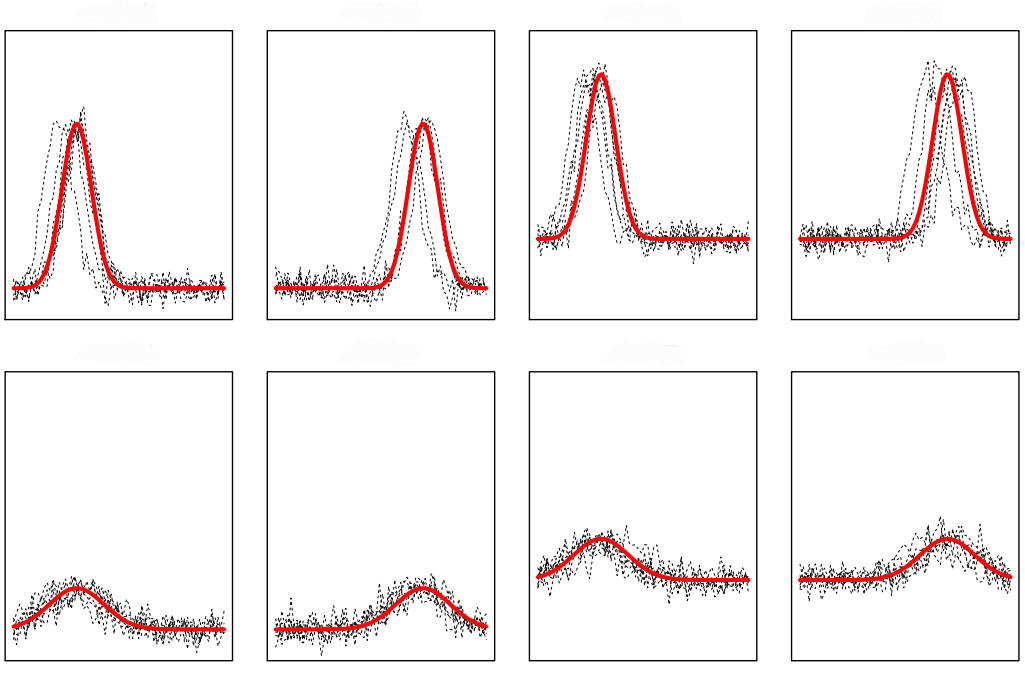
  - $m(\cdot)$ block-specific mean shape function
  - $\alpha_{ij}^{kl} = (\alpha_{ij,1}^{kl}, \alpha_{ij,2}^{kl}, \alpha_{ij,3}^{kl}) \sim \mathcal{N}(\mu_{kl}^{\alpha}, \Sigma_{kl}^{\alpha})$ vector of cell and block-specific random parameters
  - $\varepsilon_{ij}(t) \sim \mathcal{N}(0, \sigma_{\varepsilon,kl}^2)$

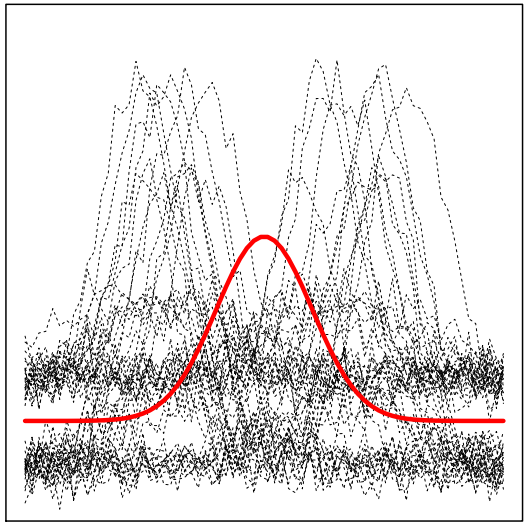
Assumption

curves in a block arise as random transformation of a block-specific mean shape function

Why SIM and co-clustering? - 1

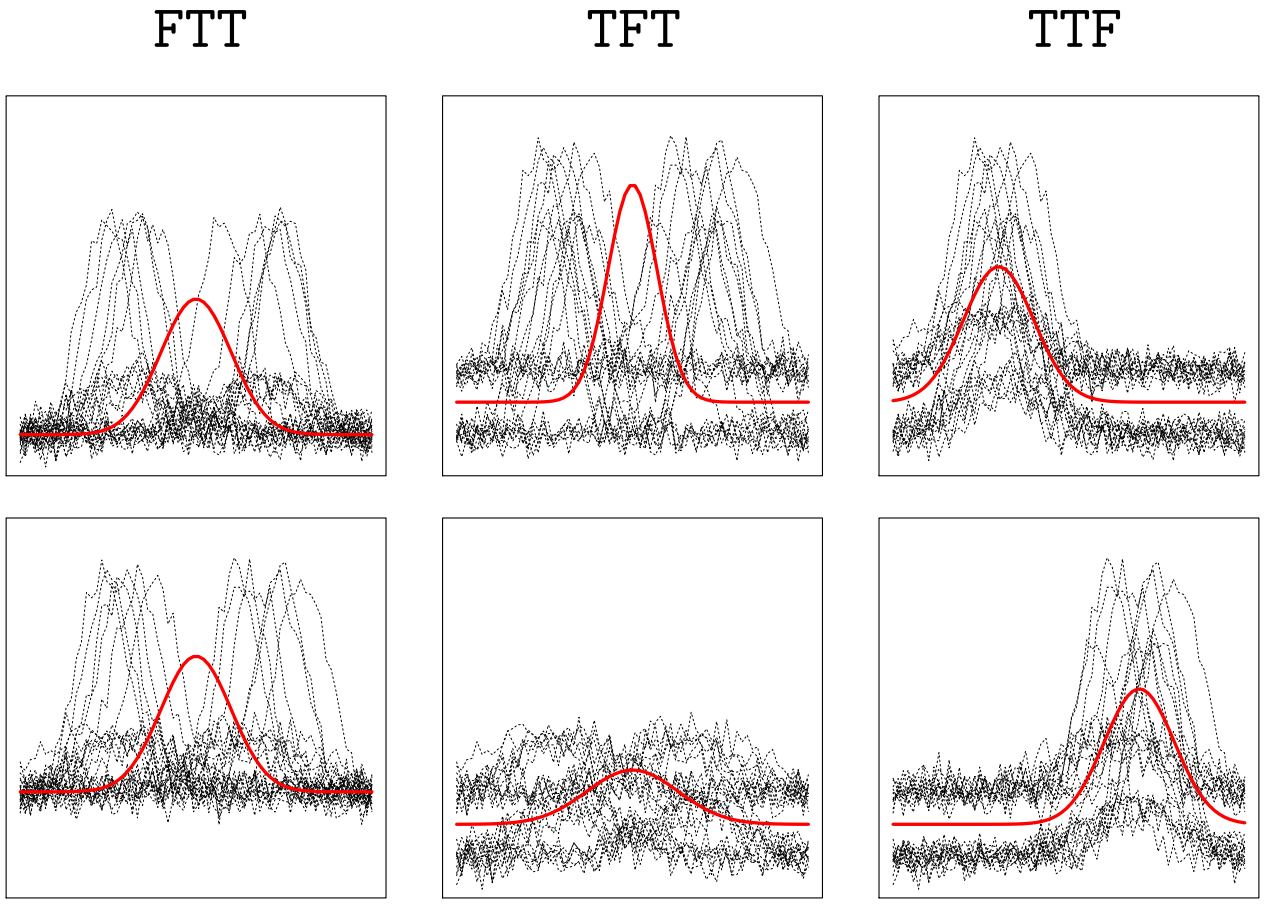
- It has been used to model both functional and longitudinal time-dependent data
- Blocks are characterized by the mean shape function but heterogeneity within them is appropriately modelled by α_{ij}^{kl}





Why SIM and co-clustering? - 2

- Further flexibility is introduced by switching off one or more random effects
  $\longrightarrow$ notion of cluster depends on subject-matter consideration



Model estimation - 1

- Maximization of the complete-data log-likelihood

$$\ell_c(\Theta, \mathbf{z}, \mathbf{w}) = \sum_{ik} z_{ik} \log \pi_k + \sum_{jl} w_{jl} \log \rho_l + \sum_{ijkl} z_{ik} w_{jl} \log p(x_{ij}; \theta_{kl})$$

- Double missing structure makes standard EM-algorithm computationally unfeasible in a co-clustering setting
  $\longrightarrow$ several modifications have been explored: CEM, SEM, VEM...
- Additional problem: no closed form expression for

$$p(x_{ij}; \theta_{kl}) = \int p(x_{ij} | \alpha_{ij}^{kl}; \theta_{kl}) p(\alpha_{ij}^{kl}; \theta_{kl}) d\alpha_{ij}^{kl}$$

since α_{ij}^{kl} enters non-linearly in the model specification

Model estimation - 2

- We propose the Marginalized SEM-Gibbs (M-SEM) algorithm
- At the h-th iteration we alternate these steps
  - Marg-step: obtain marginal distribution via MC integration

$$p(x_{ij}; \theta_{kl}^{(h-1)}) \simeq \frac{1}{M} \sum_{m=1}^{M} p(x_{ij} | \alpha_{ij}^{kl,(m)}; \theta_{kl}^{(h-1)}) \,,$$

    with $\alpha_{ij}^{kl,(1)}, \ldots, \alpha_{ij}^{kl,(M)}$ drawn from $\mathcal{N}(\mu_{kl}^{\alpha,(h-1)}, \Sigma_{kl}^{\alpha,(h-1)})$

  - SE-step: generate $(\mathbf{z}^{(h)}, \mathbf{w}^{(h)})$ via Gibbs sampler
  - M-step: Estimate $\hat{\Theta}^{(h)}$ conditionally on $(\mathbf{z}^{(h)}, \mathbf{w}^{(h)})$
    - Mixture proportions updated as usual
    - $\theta_{kl}^{(h)} = (\mu_{kl}^{\alpha,(h)}, \Sigma_{kl}^{\alpha,(h)}, \sigma_{\epsilon,kl}^{2,(h)}, \beta_{kl}^{(h)})$ estimated via approximate maximum likelihood approach for `nlme`

Model selection

- Need to select K and L, the number of row and column clusters + the random effect configuration (`FFF`, `TFF`, `TTF`...)
- In this framework we consider the ICL-BIC criterion

$$ICL = \ell_c(\hat{\Theta}, \hat{z}, \hat{w}) - \frac{K-1}{2}\log n - \frac{L-1}{2}\log d - \frac{KL\nu}{2}\log nd$$

- Random effects and time-dependent data could make model selection more troublesome $\longrightarrow$ bias towards overestimation

Incorporating prior knowledge and subject-matter considerations is highly beneficial

Some results - Pollen data

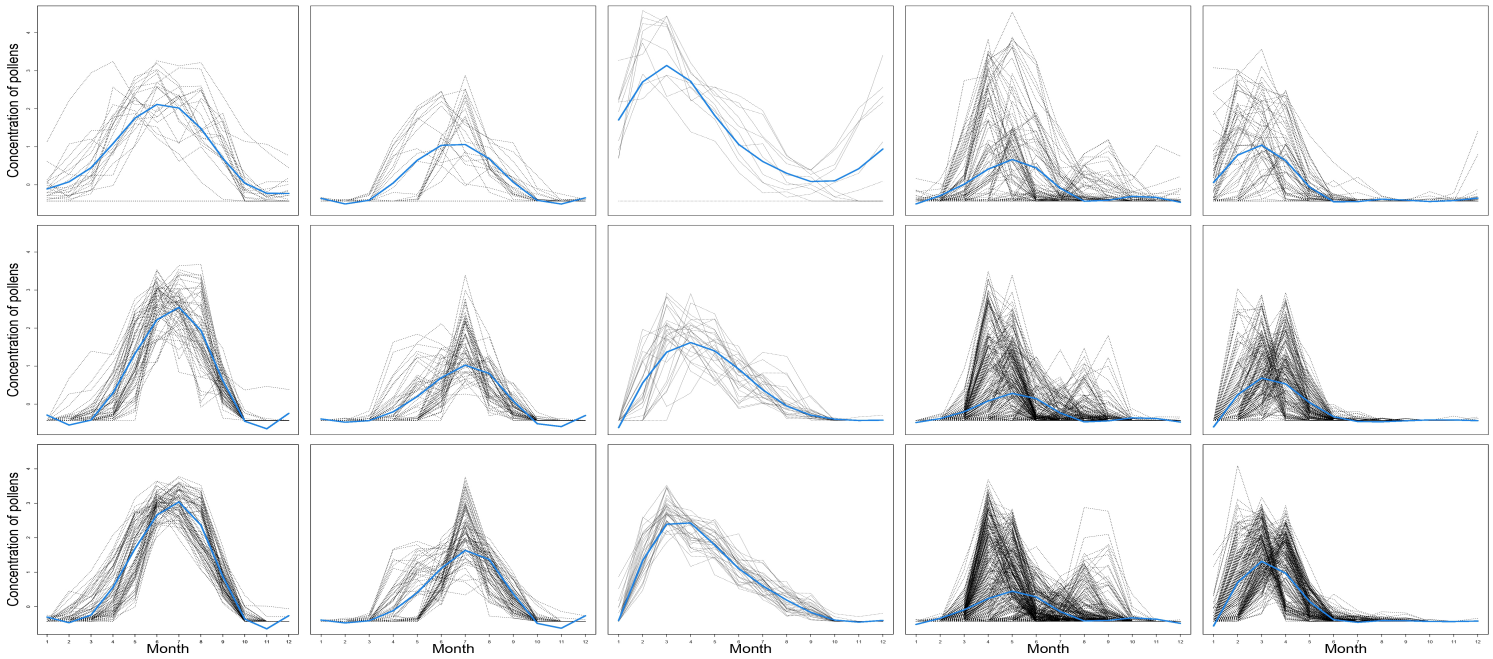
- Monthly concentration of pollens in French cities in 2016
- $p = 21$ pollens for $n = 71$ cities over $T = 12$ months
- Aim: identify homogeneous trends in pollen concentration over the year and across different geographic areas
  $\longrightarrow$ partition of both cities and pollens

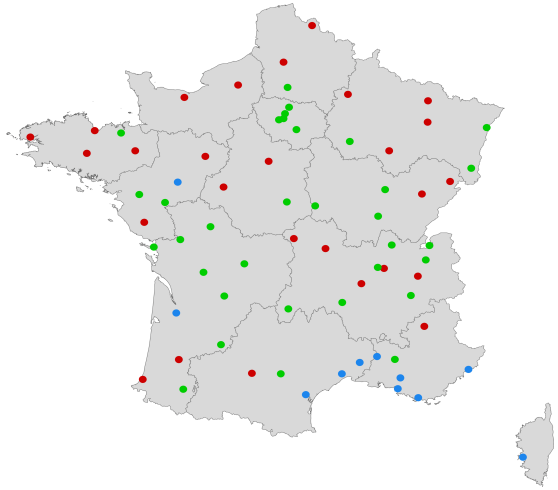
- Searching for groups of pollens differentiating for either the period of exhibition or the time span they are present

  $\downarrow$

  Only models with y-axis shift (TFF) have been estimated, $\alpha_{ij,2}$ and $\alpha_{ij,3}$ are switched off

Some results - Pollen data





- Proper discrimination of pollens according to their seasonality
  + distinguish tree pollens from weed and grass ones
- Highlight a Mediterranean region

Concluding remarks

- The proposed method partitions three-way matrices in blocks of homogeneous curves
- Some relevant advantages with respect to the competitors
  - Interpretability of the results
  - Higher flexibility, different notions of cluster
  - Both longitudinal and functional data

- Directions and open questions
  - Possible alternative model selection approaches
  - Different specification for the mean shape function
  - Bayesian estimation strategies

Some references

Casa, A., Bouveyron, C., Erosheva, E. & Menardi, G. (2021).
Co-clustering of time-dependent data via the Shape Invariant Model.
J Classif, doi.org/10.1007/s00357-021-09402-8

Other relevant references

- Bouveyron, C., Bozzi, L., Jacques, J. & Jollois, F.X. (2018). The functional latent block model for the co-clustering of electricity consumption curves.
  J R Stat Soc C, 67(4), 897-915.
- Lindstrom, M.J. (1995). Self-modelling with random shift and scale parameters and a free-knot spline shape function. *Stat Med*, 14(18), 2009-2021.
- Pinheiro, J. & Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J Comput Graph Stat*, 4(1), 12-35.
- Telesca, D. & Inoue, L.Y.T. (2008). Bayesian hierarchical curve registration.
  J Am Stat Assoc, 103(481), 328-339.